Comparaisons textométriques de traductions franco-chinoises

[Traductions franco-chinoises]

Jun MIAO, André SALEM[1]

silaomiaomiao@yahoo.fr, salem@msh-paris.fr

Résumé : Après un bref rappel sur le système d'écriture chinoise et ses prises en charge par différents systèmes de codage informatique (§3), on compare les dépouillements textométriques d'un texte français et d'une de ses traductions chinoises. Après le dépouillement du texte chinois en caractères isolés (§4), on compare un dépouillement automatisé en mots de ce même texte avec le texte français original (§5). La dernière section est consacrée à l'étude des perspectives ouvertes par la démarche textométrique pour l'analyse des différentes traductions chinoises utilisées pour rendre un même mot français (§6).

Mots-clés : textométrie; caractères chinois (hanzi); littérature; traductologie.

Abstract : After a short recall of the Chinese writing system and on its various encoding systems (§3), the authors apply textometric methods to compare a French text with its Chinese translation. After an examination of the Chinese text with isolated characters (§4), the same text cut into words with a Chinese word separation program is compared with the French original (§5). The last section aims at studying the perspectives of textometric approaches in the analysis of different Chinese translations of French words (§6).

Key-words: Textometry, Chinese characters (Hanzi), literature, translation studies.

摘要:中国文字（书写）系统是基于汉字的一种古老语言誊写方式。在对此文字系统以及其现代各种信息编码作一简单描述后(§3) 作者运用词量法对法语著作及其中文翻译进行了比较与分析。首先运用中间加空隔的方式将中文翻译当中的汉字相互独立开来(§4),其次运用分词软件对这同一翻译进行了单词的自动切分并加与原文做对比(§5)。文章的最后一部分侧重于运用词量法对法语单词的各种汉语翻译的考察(§6)。

关键词：词量法; 中文（汉字）; 文学；翻译学

1 Contexte de la recherche

Parmi les nombreuses raisons qui peuvent expliquer le fait que les méthodes d'analyse des textes sur ordinateur, de plus en plus largement répandues dans le monde occidental, ne se sont développées que plus tardivement dans la sphère culturelle chinoise, on doit considérer les facteurs liés à l'existence d'un système d'écriture très ancien, dont certaines qualités sont indiscutables, mais dont l'informatisation s'est révélée beaucoup plus complexe que celle des systèmes basés sur l'utilisation d'un alphabet réduit.

Dans la période récente, parallèlement aux efforts entrepris par les linguistes chinois pour simplifier la représentation des caractères *hanzi*, les problèmes liés à l'informatisation des systèmes d'écritures complexes ont été dépassés par la mise en place de normes internationales (telle la norme *Unicode*) et de technologies permettant la saisie et l'affichage

[1] Les auteurs remercient Kim Gerdes, Serge Fleury et Cédric Lamalle pour leur aide et leurs conseils précieux dans la réalisation de ce travail.

de textes écrits dans des langues jusqu'alors difficilement accessibles au traitement sur ordinateur.

Ces avancées technologiques ouvrent la voie à un formidable développement des études textuelles appliquées à des gisements textuels que les codages traditionnels étaient incapables de prendre en charge. Au delà de l'exploration des corpus électroniques à des fins de recherches linguistiques ou sociolinguistiques, la fouille de données textuelles concerne dorénavant un très vaste ensemble de textes saisis dans le cadre d'activités entreprises dans tous les secteurs de la vie socio-économique d'un pays en plein développement.

L'étude de bitextes, dont l'un des volets est constitué par la traduction de l'autre, constitue une entrée privilégiée dans le domaine des études comparatives entre textes rédigés dans des langues différentes. Dans ce cas, en effet, les caractéristiques quantitatives calculées à partir de chacun des volets du corpus peuvent être directement utilisées pour cerner les différences entre les langues mises en présence. C'est ce que nous allons tenter de faire dans l'étude qui suit afin de poser les premiers jalons d'études traductologiques que nous nous proposons d'entreprendre par la suite.

2 Le système d'écriture chinois

Les écritures chinoise, japonaise et coréenne utilisent, toutes trois, les caractères ***Han***, caractères d'origine chinoise dits 汉字 *(hanzi)* en chinois, ainsi que des caractères nationaux propres à chacune des langues[2]. Le chinois possède, pour sa part, un système d'écriture qui n'est ni alphabétique, ni phonétique. On peut dire que chaque caractère correspond plus ou moins à un morphème et à une syllabe de l'oral.

Le nombre de *hanzis* différents utilisés par ces systèmes d'écriture se compte en milliers (parfois en dizaines de milliers) dépassant de très loin le nombres des lettres qui permettent de transcrire les écritures alphabétiques. On dit que pour lire un journal, un lecteur chinois doit pouvoir identifier sans mal 5 000 *hanzis* environ.

2.1 Les caractères chinois

Chaque caractère chinois est composé d'un certain nombre de *traits* que l'on peut retrouver dans une série d'autres caractères. Les caractères correspondent à la fois à un segment sonore, la syllabe, et à une unité de sens[3].

人 - rén, *homme*; 大 (一+人) - dà, *grand*; 天 (二 +人) - tiān, *ciel*.

木 - mù, *bois*; 林 (木+木) - lín, *fôret*; 森 (木+林)- sēng, *grande forêt*.

Chaque caractère véhicule une signification, mais ne constitue pas nécessairement à lui seul un mot. Certains caractères changent de sens dans la combinaison avec d'autres.

东 - dōng, *Est*, 西 - xī, *Ouest*; 东西 – dōngxī, *chose*.

[2] Le Consortium Unicode et l'ISO considèrent que les caractères chinois, coréens et japonais sont les mêmes, que seuls les *glyphes* diffèrent. On peut rapprocher cette différence d'aspect des traditions différentes qui ont longtemps prévalu en allemand (police de caractères gothique), en français (police à sérifs) et en anglais (police sans sérifs). Les caractères sont codés de la même façon. Chaque tradition utilise une police appropriée pour afficher les caractères dans le style qui convient le mieux aux habitudes locales.
Après l'établissement de la République Populaire de Chine en 1949, les autorités ont entrepris des efforts pour simplifier les caractères chinois. En 1955, le Comité pour la Réforme de l'Écriture (*Wenzi gaige wei yuanhui*) a publié une proposition de caractères simplifiés. En 1964, il a publié une deuxième liste de simplifications. Cette dernière liste règle actuellement l'emploi des caractères chinois.

[3] Cf. , par exemple, [ALLETON 1997],. p.11-18.

2.2 Les mots chinois

C’est la combinaison de deux caractères ou parfois de trois caractères qui constitue le mot.

你 nǐ, *tu, toi*

好 hǎo, *bon, bien*

你好！ nǐhǎo*! Bonjour! Comment ça va?*

Dans la langue moderne, il existe beaucoup de mots bi-syllabiques, voire tri-syllabiques. Par suite de l’évolution de la langue et de l’adoption de mots empruntés à d’autres langues. Par exemple :

(1)	(2)	(3)	(4)
手，	手机	邂逅	巧克力
shǒu	shǒu jī	xiě hòu	qiǎo kė lì
main	portable	rencontre par hasard	chocolat

Dans le premier exemple, le caractère 手 (shǒu) signifie *main,* il constitue une syllabe et correspond en même temps à un sens indépendant. Dans ce cas, il peut être considéré comme un mot.

Dans le deuxième exemple, 手机, le même caractère est associé au caractère 机 (jī, *machine, appareil*) Il garde dans ce cas le sens *main*, mais la combinaison des deux caractères prend un nouveau sens : *téléphone mobile, portable*.

Dans le troisième exemple, la combinaison des deux caractères 邂逅 (xiè hòu) signifie *se rencontrer par hasard*, mais ces caractères perdent leur sens lorsqu’il sont isolés.

Dans le mot 巧克力 (anglais *chocolate*), chacun des caractères 巧, 克, 力 possède un sens propre sans rapport immédiat avec le mot (巧: *adroite*, *habile*; 克: *convaincre*; 力: *force*).

Produit courant, 茉莉花茶 (mò lì huā chā, *le thé au jasmin*) est un mot, dont les composants identifiables sont difficiles à segmenter. On peut considérer 茉莉 (mò lì, *jasmin*) comme un mot bi-syllabique composé de deux caractères dépourvus de sens propre. Mais en combinaison avec le caractère 花 (huā, *fleur*), le mot qui désigne toujours le jasmin, renvoie à la fleur de l’arbuste. On peut considérer le caractère 茶 (chā, *thé*) comme un mot monosyllabique. Mais précédé par le caractère 花 (huā, *fleur*), on peut également considérer que les caractères combinés 花茶 (huā chā, *thé aux fleurs*) qui sont différents de 绿茶 (lǜ chá, *thé vert*) ou de 红茶 (hóng chá, *thé noir*) forment un nouveau mot.

2.3 Les phrases et la ponctuation

Comme dans le cas des mots, il est difficile de définir clairement les limites de la phrase chinoise. Les définitions et les classifications de la phrase que l’on trouve dans les grammaires chinoises (phrases énonciatives, interrogatives, impératives, exclamatives, etc.) permettent difficilement de segmenter un texte en phrases de manière automatisée.

La ponctuation est d’usage récent en chinois. En 1919, on a commencé à utiliser la ponctuation moderne en se référant au système de ponctuation occidental. Le système utilisé actuellement conserve la trace des réformes successives de l'écriture chinoise. C'est pourquoi

la ponctuation chinoise moderne, malgré ses similarités avec celle utilisée en occident, reste distincte de cette dernière.

L'utilité des repères liés à la notation de la ponctuation chinoise est d'autant plus importante que, comme on s'en souvient, les mots (ou plutôt les caractères) chinois sont écrits l'un après l'autre sans être séparés par des espaces. [4]

3 Le codage informatique des caractères chinois

En raison de leur nombre élevé et contrairement à ce qui se passe pour les systèmes d'écriture des langues qui utilisent un alphabet restreint, les caractères chinois ne peuvent être représentés à l'aide d'un codage sur un seul octet. La norme *Unicode* qui permet de représenter chaque caractère sur plusieurs octets fournit une bonne solution pour représenter les caractères chinois.[5].

3.1 Logiciels supportant le traitement de textes chinois.

Dans leurs versions récentes, les logiciels de traitement de textes permettent de manipuler, en plus des textes codés en unicode qui vont rapidement constituer la norme, des polices multioctets qui permettent d'afficher correctement les textes chinois (entre autres écritures non latines). Avec le logiciel Word[6], par exemple, lorsqu'on tente d'enregistrer un texte chinois, avec l'option texte seulement une boite de dialogue permet de sélectionner le codage `Chinois simplifié (GB2312)` comme on peut le voir sur la figure 1.

3.2 Lexico3 et les textes chinois

Dans ses versions actuelles (3.5.0.2), ***Lexico3*** manipule des chaînes de caractères codés sur un seul octet. Cette limite, qui est en voie d'être dépassée[7], n'entraîne cependant pas l'impossibilité de traiter des chaînes de caractères codées sur plusieurs octets. Comme on comprend, en les comparant octet par octet, il est possible de conclure que deux chaînes de caractères multioctets sont identiques ou qu'elles sont différentes. De plus, les systèmes informatiques modernes permettent d'afficher correctement certaines représentation multioctets qui ne sont pas des représentations unicode .

Pour le présent travail, nous avons utilisé un codage **`Chinois simplifié . Mainland China`** proposé par le logiciel ***Word***. On prend en charge ce codage sous Lexico3 en activant l'article `Chinois simplifié.Mainland China` proposé par le menu `Options` (couteau suisse) de ***Lexico3***.

Les composants utilisés dans ***Lexico3*** (Edition du texte, Concordances, Carte des sections, etc.) affichent ce codage correctement lorsqu'on choisit de le visualiser avec le codage `Chinois GB2313` des navigateurs :

```
Bouton droit -> Codage -> Plus ->. Chinois simplifié (GB2312)
```

---

4 A l'instar de très nombreux systèmes d'écriture parmi lesquels ceux de l'antiquité (latin, grec, hébreu, sumérien, etc.).

5 Un grand nombre de systèmes d'écriture occidentaux, dont le système du français ont utilisé jusqu'à une date récente le code ASCII (127 caractères), puis le code ASCII étendu (255 caractères) qui permettait de coder en outre les voyelles accentuées du français.

6 Nous avons utilisé, pour cette étude, la version 2003 du logiciel ***Word*** distribué par Microsoft.

7 Plusieurs versions de la série *Lexico*, en cours d'achèvement, permettent déjà de traiter les chaînes de caractères unicodes. Le logiciel *MKAlign*, développé par S. Fleury dans l'équipe Syled-Cla2t permet également de traiter les textes encodés sous ces formats.

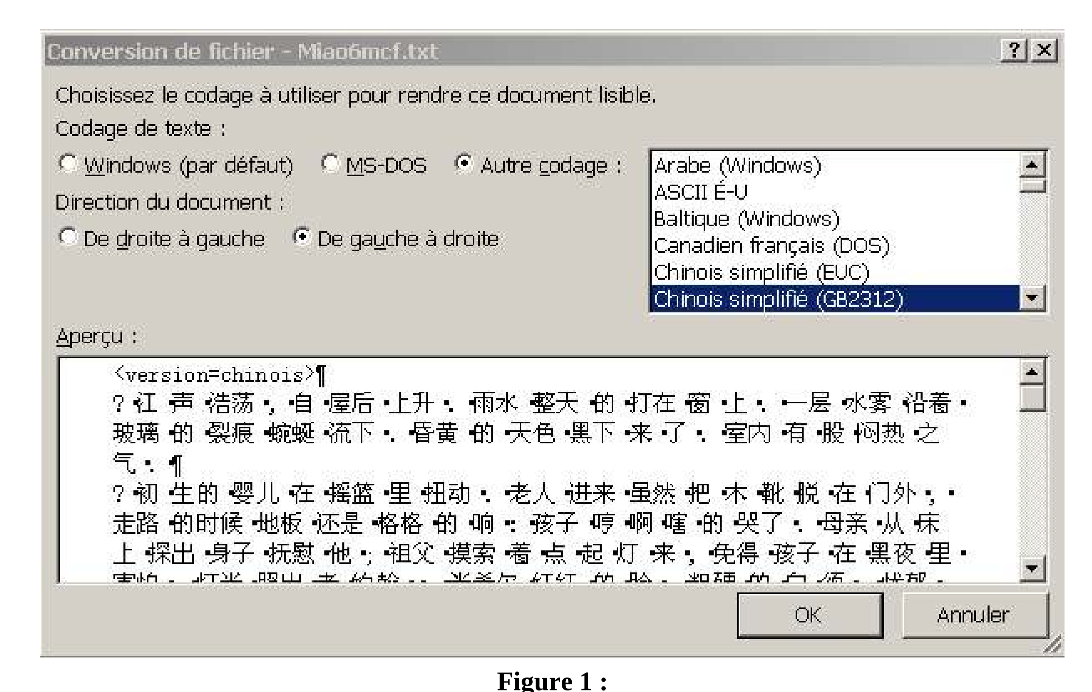


Figure 1 :

Word 2003 : Paramétrage de l'enregistrement du texte

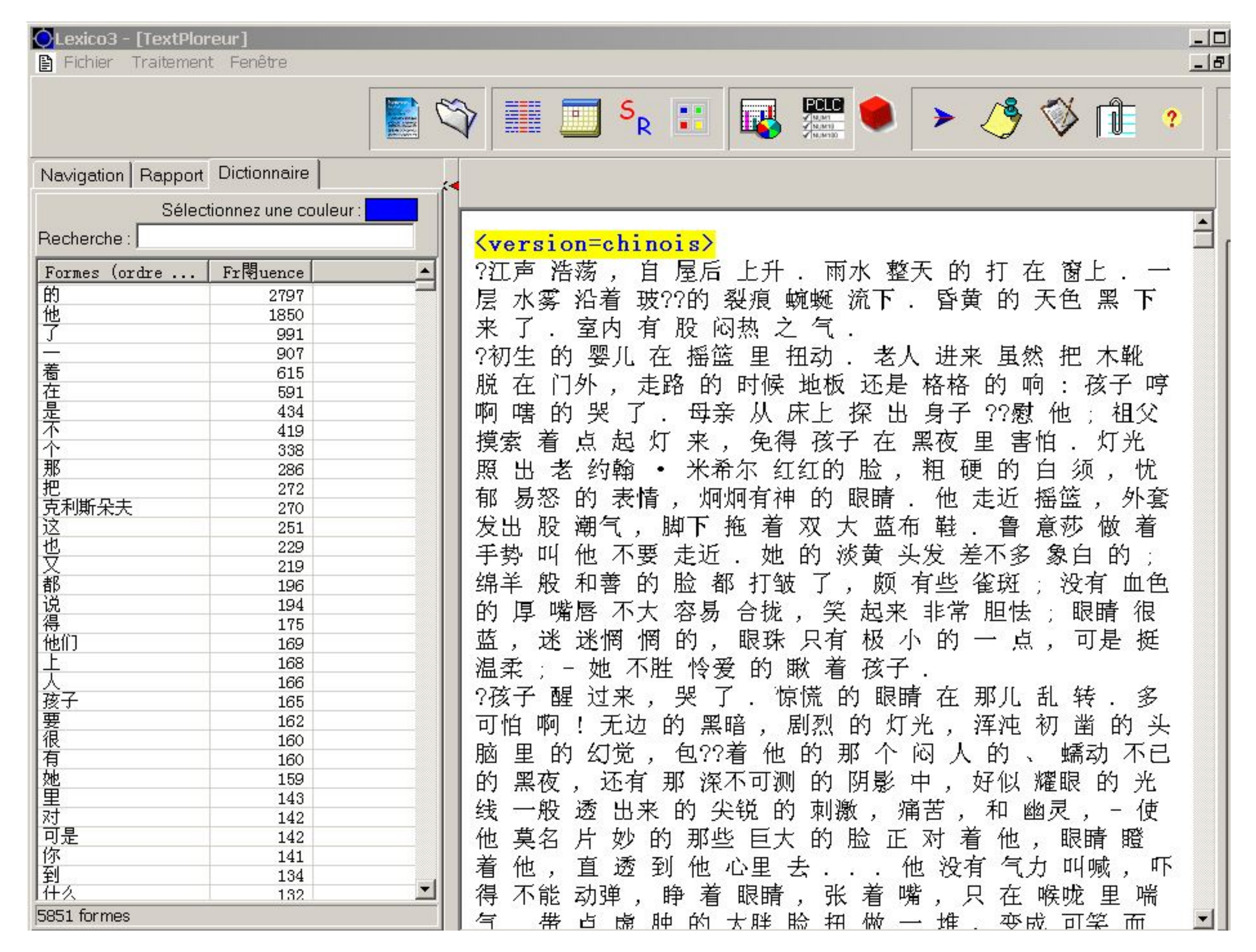


Figure 2 :

Lexico 3 : Affichage du texte avec le codage « Chinois simplifié (GB2312) »



Figure 3 :
Lexico 3 : Fenêtre de réglage du paramètre « encodage des caractères »

4 Un corpus d'application

Pour illustrer ces possibilités, nous utiliserons un corpus bilingue dont le volet français est constitué par le premier chapitre du roman *Jean-Christophe* publié en 1904 par Romain Rolland (1866-1944). On trouve, au tableau 1, ci-dessous un extrait du texte original de Romain Rolland. Le second volet du corpus est constitué par la traduction de ce texte en chinois par Fu Lei (1908-1966). Nous appellerons respectivement ces deux corpus ***JC1-Français*** et ***JC1-Chinois.***

4.1 Segmentation du texte en caractères

Comme on l'a vu plus haut, sans que cela constitue une gêne pour le lecteur expérimenté, le système d'écriture chinois n'utilise pas d'espace entre les unités lexicales placées côte à côte. Cette circonstance constitue une difficulté spécifique pour l'exploitation textométrique des textes chinois.

Sur quels critères peut-on s'appuyer pour découper des unités statistiques au fil du texte afin de réaliser des comparaisons entre textes ? Pour cette première analyse, nous nous appuierons sur une segmentation automatique, relativement facile formaliser et à mettre en œuvre sur un ordinateur, qui isole chaque caractère *hanzi*.[8]

[8] Pour réaliser cette segmentation en caractères, nous avons remplacé, en utilisant pour cela une expression régulière, chaque caractère du texte de départ par ce même caractère précédé d'un espace (code ASCII 32). Le fichier ainsi modifié réalise l'isolation de tous les caractères du corpus. Une procédure de ce type est disponible à l'adresse : http://www.cavi.univ-paris3.fr/

Tableau 1

Extrait des corpus ***JC1-Fr*** et ***JC1-Chin***,

Le grondement du fleuve monte derrière la maison. La pluie bat les carreaux depuis le commencement du jour. Une buée d'eau ruisselle sur la vitre au coin fêlé. Le jour jaunâtre s'éteint. Il fait tiède et fade dans la chambre.

Le nouveau-né s'agite dans son berceau. Bien que le vieux ait laissé, pour entrer, ses sabots à la porte, son pas a fait craquer le plancher : l'enfant commence à geindre. La mère se penche hors de son lit, afin de le rassurer ; et le grand-père allume la lampe en tâtonnant, pour que le petit n'ait pas peur de la nuit. La flamme éclaire la figure rouge du vieux Jean-Michel, sa barbe blanche et rude, son air bourru et ses yeux vifs. Il vient près du berceau. Son manteau sent le mouillé ; il traîne en marchant ses gros chaussons bleus. Louisa lui fait signe de ne pas s'approcher. Elle est d'un blond presque blanc ; ses traits sont tirés ; sa douce figure mouton est marquée de taches de rousseur ; elle a des lèvres pâles et grosses, qui ne parviennent pas à se rejoindre et qui sourient avec timidité ; elle couve l'enfant des yeux – des yeux très bleus, très vagues, où la prunelle est un point tout petit, mais infiniment tendre.

§ l'enfant s'éveille et pleure. son regard trouble s'agite. quelle épouvante ! les ténèbres, l'éclat brutal de la lampe, les hallucinations d'un cerveau à peine dégagé du chaos, la nuit étouffante et grouillante qui l'entoure, l'ombre sans fond d'où se détachent, comme des jets aveuglants de lumière, des sensations aiguës, des douleurs, des fantômes : ces figures énormes qui se penchent sur lui, ces yeux qui le pénètrent, qui s'enfoncent en lui, et qu'il ne comprend pas - il n'a pas la force de crier ; la terreur le cloue immobile, les yeux, la bouche ouverts, soufflant du fond de la gorge. sa grosse tête boursouflée se plisse de grimaces lamentables et grotesques ; la peau de sa figure et de ses mains est brune, violacée, avec des taches jaunâtres.

Romain Rolland, *Jean-Christophe,* 1904

第一部

江声浩荡,自屋后上升.雨水整天的打在窗上.一层水雾沿着玻璃的裂痕蜿蜒流下.昏黄的天色黑下来了.室内有股闷热之气.

初生的婴儿在摇篮里扭动.老人进来虽然把木靴脱在门外,走路的时候地板还是格格的响:孩子哼啊嗐的哭了.母亲从床上探出身子抚慰他;祖父摸索着点起灯来,免得孩子在黑夜里害怕.灯光照出老约翰·米希尔红红的脸,粗硬的白须,忧郁易怒的表情,炯炯有神的眼睛.他走近摇篮,外套发出股潮气,脚下拖着双大蓝布鞋.鲁意莎做着手势叫他不要走近.她的淡黄头发差不多象白的;绵羊般和善的脸都打皱了,颇有些雀斑;没有血色的厚嘴唇不大容易合拢,笑起来非常胆怯;眼睛很蓝,迷迷惘惘的,眼珠只有极小的一点,可是挺温柔;-她不胜怜爱的瞅着孩子.

孩子醒过来,哭了.惊慌的眼睛在那儿乱转.多可怕啊!无边的黑暗,剧烈的灯光,浑沌初凿的头脑里的幻觉,包围着他的那个闷人的*蠕动不已的黑夜,还有那深不可测的阴影中,好似耀眼的光线一般透出来的尖锐的刺激,痛苦,和幽灵,使他莫名片妙的那些巨大的脸正对着他,眼睛瞪着他,直透到他心里去...他没有气力叫喊,吓得不能动弹,睁着眼睛,张着嘴,只在喉咙里喘气.带点虚肿的大胖脸扭做一堆,变成可笑而又可怜的怪样子;脸上与手上的皮肤是棕色的,暗红的,还有些黄黄的斑点.

Traduction chinoise par Fu Lei, 1957[9]

[9] Nous avons utilisé la version complète, réunie en 1957 par les Éditions Littéraires Populaires (人民文艺出版社), à partir d'une révision par Fu Lei de la première version de 1953.

Le tableau 1 montre un extrait du texte original suivi de sa traduction chinoise.
La figure 4 montre, dans la fenêtre de droite, l'affichage par ***Lexico 3*** du texte chinois dans lequel les caractères ont été isolés par insertion d'un caractère espace entre chaque caractère. Dans la fenêtre de gauche on peut lire le résultat du dépouillement statistique réalisé sur la base du décompte des caractères isolés. Les caractères sont triés par ordre de fréquence décroissante dans le corpus analysé.

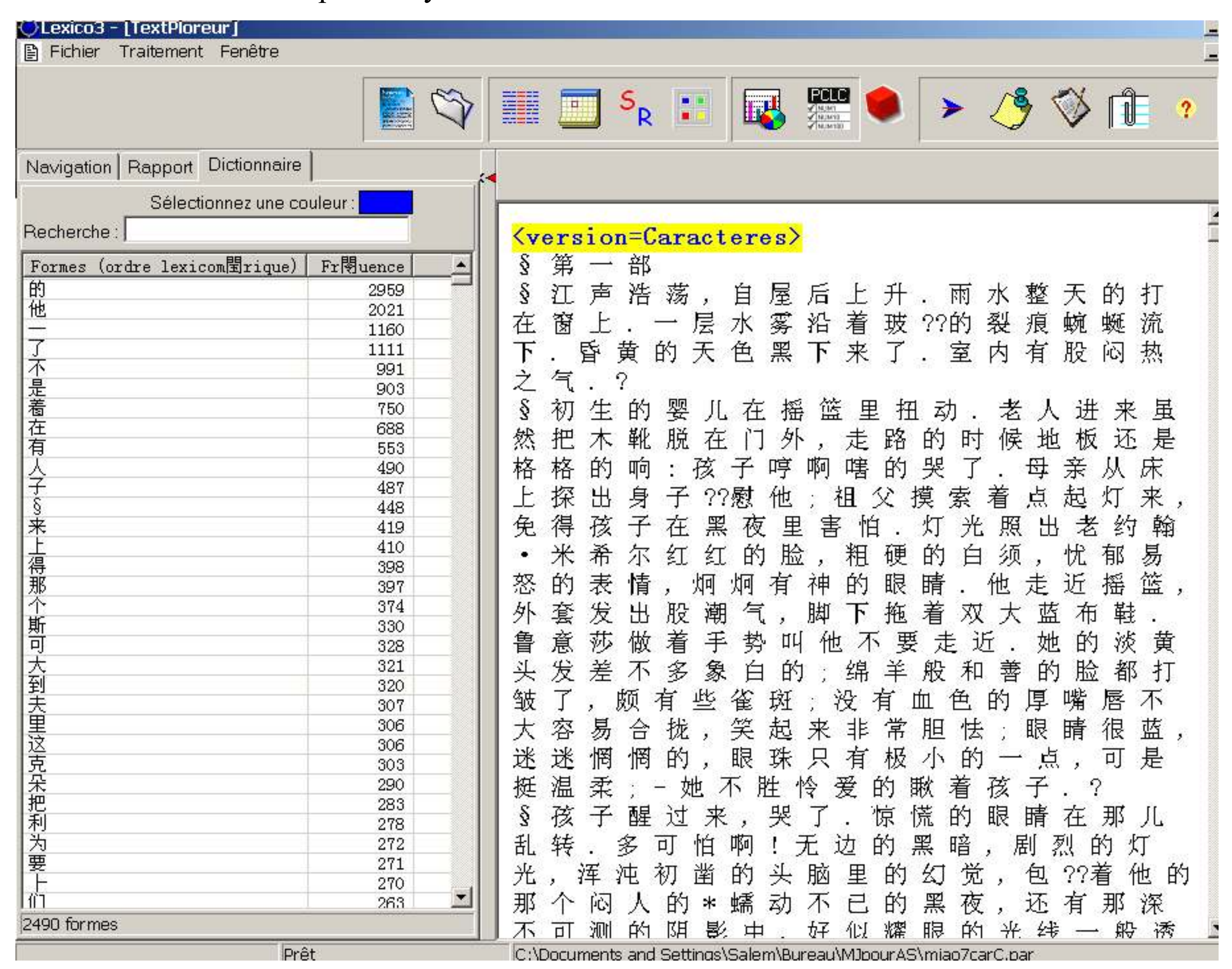


Figure 4

Exploitation avec ***Lexico3*** du texte chinois découpé en caractères

Le texte ainsi modifié va nous permettre d'obtenir un premier dépouillement en caractères (*hanzi*) du volet chinois du corpus. On peut voir les principales caractéristiques quantitatives de ce dépouillement au tableau 2.

Tableau 2

Principales caractéristiques quantitatives résultant du dépouillement en caractères (*hanzi*) du volet chinois du corpus

Partie	NB de caract.	Caract. différents	hapax	FMax	
Caractères	56 797	2 478	579	2 959	的

Le tableau 2 montre que les 56 797 caractères que compte le corpus ***JC1-Chin*** sont des occurrences de 2 478 hanzis différents. Un quart environ de ces caractères, soit 579, ne

trouvent qu'une seule occurrence dans le corpus. Le caractère le plus fréquent est le caractère 的 (qui correspond plus ou moins à la préposition *de* en français).

La figure 5, qui rend compte de l'apparition de nouveaux caractères au fur et à mesure que l'on parcourt le texte, permet de préciser ces observations. La courbe d'accroissement réalisée à partir des caractères *hanzis* montre qu'on atteint, dès les 5 000 premiers caractères du texte le seuil de 1 000 caractères différents. Les 5 000 caractères suivants n'apportent que 500 nouveaux *hanzis*. Comme dans le cas des courbes d'accroissement de vocabulaire constituées à partir des mots, les tranches successives apportent de moins en moins d'unités nouvelles. Dans le cas des *hanzis* cependant on peut remarquer que l'accroissement initial est plus fort que dans le cas des courbe d'accroissement réalisées à partir d'unités lexicales (cf. § 5, infra).



Figure 5

Apparition progressive des caractères dans le volet chinois.

4.2 Segmentation automatique en « mots »

Certains professionnels du Traitement Automatique des Langues proposent sur le web des procédures qui permettent de découper un texte chinois en « mots ». Dans cette section, nous utiliserons un découpage automatique en mots réalisé par un logiciel de segmentation spécialement conçu pour les textes chinois[10]. On peut voir au tableau 3 le résultat de cette segmentation en mots réalisée à partir de l'extrait de texte présenté au tableau 1.

---

[10] Pour cette première étude, nous avons utilisé le logiciel 海量智能分词研究版 (*Hailanda Segmentation intelligente* - version d'essai) réalisé par le Centre d'intelligence artificielle *Hailanda*, disponible à l'adresse suivante : http://www.mydown.com/code/234/234301.html . En plus de la segmentation, ce logiciel réalise une catégorisation des mots du texte orientée vers la recherche d'information technico-commerciale. Nous n'avons pas utilisé cette catégorisation pour notre étude. Il existe d'autres logiciels de segmentation du chinois, que l'on peut trouver sur l'Internet : ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System), et

Tableau 3

Extrait du volet chinois ***JC1-Chin*** segmenté en mots
(Chaque mot isolé par le logiciel *Hailanda* est suivi d'un blanc)

第一 部

江 声 浩荡 , 自 屋后 上升 . 雨水 整天 的 打在 窗 上 . 一层 水雾 沿着 玻璃 的 裂痕 蜿蜒 流下 . 昏黄 的 天色 黑下 来 了 . 室内 有 股 闷热 之气 . 初 生的 婴儿 在 摇篮 里 扭动 . 老人 进来 虽然 把 木 靴 脱 在 门外 , 走路 的时候 地板 还是 格格 的 响 : 孩子 哼 啊 嗜 的 哭了 . 母亲 从 床上 探出 身子 抚慰 他 ; 祖父 摸索 着 点 起 灯 来 , 免得 孩子 在 黑夜 里 害怕 . 灯光 照出 老 约翰 · 米希尔 红红 的 脸 , 粗硬 的 白 须 , 忧郁 易怒 的 表情 , 炯炯有神 的 眼睛 . 他 走近 摇篮 , 外套 发出 股 潮气 , 脚下 拖着 双 大 蓝布 鞋 . 鲁 意 莎 做着 手势 叫 他 不要 走近 . 她的 淡黄 头发 差不多 象 白的 ; 绵羊 般 和善 的 脸 都 打 皱 了 , 颇有 些 雀斑 ; 没有 血色 的 厚 嘴唇 不大 容易 合拢 , 笑 起来 非常 胆怯 ; 眼睛 很 蓝 , 迷 迷惘 惘 的 , 眼珠 只有 极小 的 一点 , 可是 挺 温柔 ; - 她 不胜 怜爱 的 瞅着 孩子 .

孩子 醒过来 , 哭了 . 惊慌 的 眼睛 在 那儿 乱转 . 多可 怕 啊 ! 无边 的 黑暗 , 剧烈 的 灯光 , 浑沌 初 凿 的 头脑 里 的 幻觉 , 包围着 他的 那个 闷人 的 * 蠕动 不已 的 黑夜 , 还有 那 深不可测 的 阴影 中 , 好似 耀眼 的 光线 一般 透出 来的 尖锐 的 刺激 , 痛苦 , 和 幽灵 , - 使 他 莫名 片 妙 的 那些 巨大 的 脸 正对着 他 , 眼睛 瞪着 他 , 直 透 到 他 心里 去 . . . 他 没有 气力 叫喊 , 吓得 不能 动弹 , 睁着 眼睛 , 张 着 嘴 , 只 在 喉咙 里 喘气 . 带 点 虚肿 的 大 胖 脸 扭 做 一堆 , 变成 可笑 而 又 可怜 的 怪样子 ; 脸上 与 手上 的 皮肤 是 棕色 的 , 暗红 的 , 还有 些 黄黄 的 斑点

5 Comparaisons quantitatives à partir des *mots*

Les comptages réalisés à partir des mots ainsi découpés par l'algorithme de segmentation permettent de comparer les résultats obtenus sur le texte chinois à ceux que l'on obtient de la même manière sur la version française du texte.

Tableau 4

Principales caractéristiques quantitatives du dépouillement en *mots*
réalisé sur les volets français ***JC1-Fr*** et chinois ***JC1-Chin*** du corpus.

Partie	Occurrences	Formes	Hapax	F. Max	
JC1-Chin	34 743	7 196	3 781	2313	的
JC1-Fr	39 666	6 673	3 970	1578	de

Comme on le voit au tableau 4, la traduction chinoise compte nettement moins de mots graphiques que le texte français. On notera qu'elle compte cependant nettement plus de

Java Lucene segmentation du chinois, etc.. Signalons qu'en raison de l'absence d'un consensus sur la définition de ce qu'est un mot chinois, aucun logiciel ne peut prétendre fournir de résultats parfaits.

formes différentes. La proportion des formes qui n'apparaissent qu'une seule fois dans chacun des textes est moindre dans le texte chinois alors que la forme la plus fréquente y trouve nettement plus d'occurrences que dans le texte français[11].

La comparaison entre le système des mots chinois et celui des caractères chinois, pour lequel nous avons présenté plus haut des décomptes comparables montre que les mots chinois sont composés en moyenne de 1,6 caractères et que le mot le plus fréquent rassemble presque toutes les occurrences du caractère le plus fréquent (dans les deux cas le caractère : 的, *de*).

Tableau 5

Les formes les plus fréquentes pour chacun des volets du corpus

	Français		Chinois	
1	de	1 578	2313	的
2	il	1 044	1581	他
3	et	1 034	638	了
4	le	908	373	在
5	la	841	368	是
6	les	575	276	夫
7	Il	515	275	朵
8	se	463	274	克利斯
9	lui	448	235	把
10	des	447	208	着
11	ne	439	204	也
12	un	407	184	他的
13	en	399	158	又
14	que	394	156	孩子
15	pas	376	147	他们
16	qui	375	143	都
17	son	362	142	可是
18	dans	329	139	来
19	une	314	139	个
			136	她

La comparaison entre les formes les plus fréquentes dans chacun des volets du corpus montre que les fréquences décroissent plus rapidement dans le volet chinois. L'étude comparée des courbes d'accroissement du vocabulaire, figure 6, précise les résultats obtenus par la comparaison des principales caractéristiques lexicométriques des volets français et chinois du corpus. La courbe située dans le haut du graphique correspond à l'enrichissement du

[11] Il nous a semblé intéressant de publier ces premiers comptages sur la comparaison textométrique entre textes chinois et textes français. Cependant, ces résultats présentés dans le but de fournir une comparaison sur deux systèmes d'écriture très différents doivent être pris avec de grandes précautions. Nous étudierons par la suite l'influence que peut avoir la lemmatisation de chacune des listes de formes sur les résultats produits de la sorte (ainsi par exemple, la fréquence de la forme chinoise la plus fréquente 的 2313 occ. renvoie à la forme française *de* 1578 occ . mais aussi aux formes *du* 243 occ., *des* 447 occ., etc.).

vocabulaire chinois au fil du texte. Le fait que ce texte comporte moins d'occurrences est responsable de l'interruption de la courbe correspondante (abscisse 34 743) avant la courbe qui correspond au texte français (abscisse 39 666). La courbe correspondant à l'apparition de nouveaux mots chinois est située, dès que l'on atteint le premier tiers du corpus, largement au-dessus de celle qui correspond à l'apparition des mots français, ce qui confirme l'existence d'un plus grand nombre de formes en chinois.

On peut remarquer que des paliers créés par le ralentissement de l'accroissement du vocabulaire au cours du récit peuvent être mis en rapport d'une courbe à l'autre. Au ralentissement qui survient sur la courbe correspondant au texte français (abscisse 20 000) correspond un ralentissent dans la traduction chinoise (abscisse 17 000). A celui qui survient pour le texte français (abscisse 32 500) correspond également un ralentissent dans la traduction chinoise (abscisse 28 000).

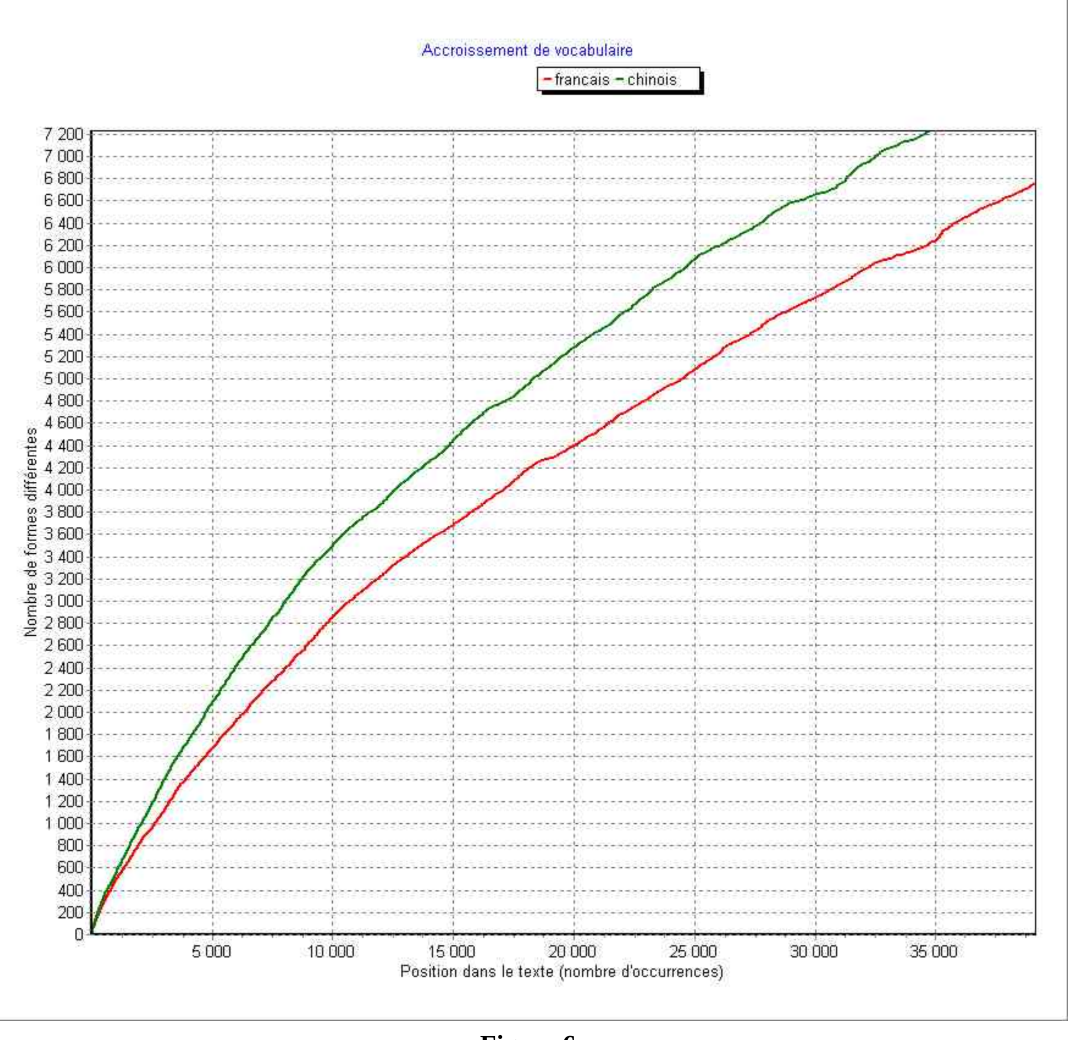


Figure 6

Courbes d'accroissement du vocabulaire réalisée
sur les volets français ***JC1-Fr*** et chinois ***JC1-Chin*** du corpus.

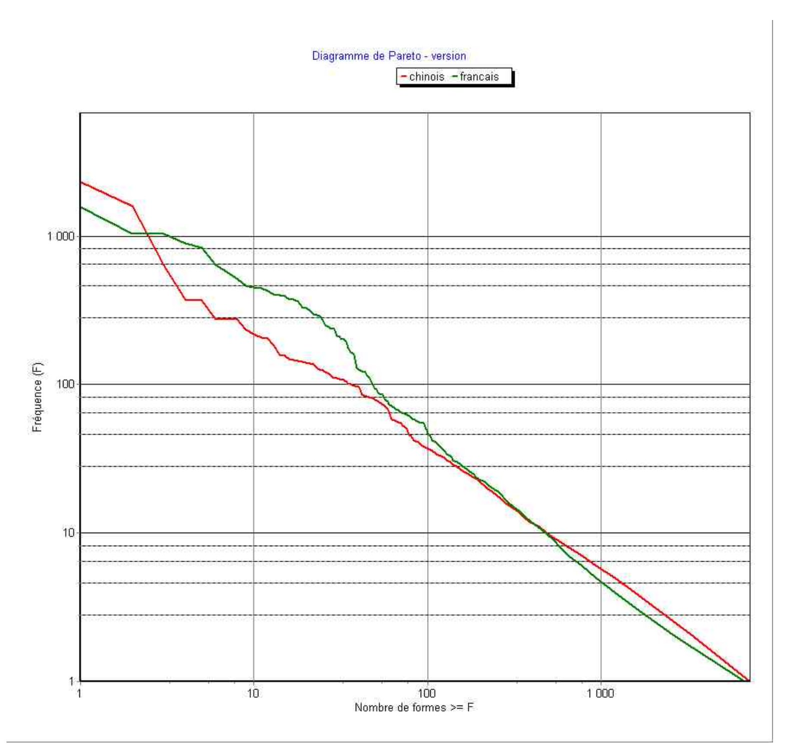


Figure 7 :

Diagramme de Pareto pour les deux volets du corpus

==== Guide de lecture pour la figure 7 ====

Pour un texte ***T*** dépouillé en unités statistiques appelées *formes*, le ***Diagramme de Pareto*** permet de visualiser la structure de la gamme des fréquences.

- ■ L'axe vertical permet de représenter la fréquence *F* des formes du textes (laquelle varie de *1* à *Fmax*, fréquence maximale calculée pour le texte ***T***).
- ■ Sur l'axe horizontal, on porte la quantité : *nombre de formes du texte dont la fréquence est supérieure à F*.
- ■ Avant de tracer le Diagramme, on transforme chacune de ces quantités en son logarithme décimal.

Le Diagramme ainsi obtenu prend alors approximativement la forme droite que l'on appelle *Droite de Zipf* en l'honneur de Georges Kingsley Zipf qui a montré que ce type de procédure réalisée à partir de larges catégories de textes permet de mettre en évidence une propriété statistique commune aux dépouillements en unités lexicales. Cette propriété est parfois présentée sous la forme excessivement simplifiée :

Rang x *Fréquence* = *Constante*

Pour en savoir plus :

Zipf, GK (1935), *The Psychobiology of Language, an introduction to Dynamic Philology*, Boston, Houghton-Mifflin.

Lebart L., Salem A., *Statistique textuelle*, Paris, Dunod, 1994, téléchargeable sur le site : http://www.cavi.univ-paris3.fr/lexicometrica/livre/st94/st94-tdm.html

La comparaison des deux courbes fait apparaître des différences assez nettes dans la structure des gammes de fréquences des deux textes. Le texte français possède nettement plus de formes dans la zone de fréquences qui s'étend de 50 occurrences à 1000 occurrences environ. De son côté, le chinois crée plus de formes différentes dans la zone des très basses fréquences.

6 Un exemple d'étude parallèle

Aligner un bitexte, c'est construire une représentation qui met en correspondance des unités textuelles en rapport de traduction mutuelle. Le tableau 6 montre un alignement des deux volets du bitexte réalisé à partir du corpus ***JC1*** au niveau du paragraphe[12].

A partir d'un tel alignement on peut s'intéresser aux traductions de ce qui constitue une unité dans la langue source dans l'autre volet du corpus. Cette comparaison peut être menée simultanément du point de vue distributionnel , à l'aide de l'outil concordance (cf. tableau 7) et d'un point de vue *spatial* (cf. figure 8).

6.2 Le groupe vieux/vieillard *et son correspondant* 老人 (lao ren)

A titre d'exemple, nous examinerons les traductions chinoises d'un ensemble de mots qui rendent en français le concept de ***vieillesse*** : *vieux, vieillard,* etc.[13] Pour cette famille de mots, nous obtenons une fréquence globale de 95 occurrences qui se répartissent comme suit :

vieux 77, vieille 7, vieil 3, vieillard 3, vieilles 2, vieillards 1, vieillissait 1, vieillots 1.

On trouve au tableau 7 un extrait de concordance réalisée autour du pôle 老(lao, *vieux*), dont les lignes sont triées par ordre d'apparition dans le texte chinois. La localisation des occurrences de chacun de ces termes dans la carte des sections établie pour le texte français (figure 8) permet de repérer des sections correspondantes du texte chinois dans lesquelles on peut s'attendre à ce que soit rendue, en chinois, l'idée de *vieux*. La liste des mots les plus spécifiques dans le texte chinois qui correspond à ces dernières sections, nous laisse penser que le concept *vieux, vieillard, etc.,* est souvent rendu en chinois par les termes 老人 (lao ren, *vieil homme*) et 老(lao, *vieux*) qui constituent par ailleurs les équivalences traductionnelles les plus adaptées pour traduire le concept de *vieux*.

Dans une seconde étape, nous introduisons les mots 老人 et 老 sur la carte des sections découpées à partir du texte chinois. La comparaison des deux volets montre que la correspondance est loin d'être parfaite. On a rassemblé dans le tableau 8 des paires, sélectionnées à partir du concept français *vieux*, qui se trouvent être en rapport de traduction avec des expressions chinoises. L'analyse des discordances dans la localisation de ces formes révèle avant tout un écart entre le champ sémantique du mot français *vieux* et celui du *hanzi* chinois 老 (lao, *vieux, ancien, etc.*). En français, le mot *vieux* possède un lien étroit avec l'âge et le temps, mais il véhicule aussi une valeur parfois péjorative lorsqu'il s'applique à des objets ou à des personnes dans certains contextes (*vieux vêtements, vieille caisse*). En chinois, tout au contraire, le mot 老, dont le champ sémantique est un peu plus large, est employé pour désigner des personnes anciennes, respectables, honorables (老师 *professeur,* 老师傅 *vieux maître*).

---

[12] Cet alignement a été réalisé en utilisant le logiciel MKAlign proposé par Serge Fleury.. ce logiciel peut être téléchargé sur le site : http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm.

[13] Maria Zimina-Poirot a étudié dans sa thèse [Zimina 2004] des correspondances traductionnelles de ce type. Les logiciels de textométrie permettent désormais l'étude systématique de ce genre de correspondances traductionnelles. Les termes de la correspondance peuvent être étendus par l'utilisation du système des expressions rationnelles. Dans notre cas, le motif : *vie[iu]* permet de localiser toutes les occurrences des formes détaillées plus haut.

Tableau 6 :

Alignement en paragraphes sur les deux volets du corpus

§ le grondement du fleuve monte derrière la maison. la pluie bat les carreaux depuis le commencement du jour. une buée d'eau ruisselle sur la vitre au coin fêlé. le jour jaunâtre s'éteint. il fait tiède et fade dans la chambre.	§ 江声 浩荡, 自 屋后 上升. 雨水 整天 的 打 在 窗上. 一 层 水雾 沿着 玻璃 的 裂痕 蜿蜒 流下. 昏黄 的 天色 黑 下来 了. 室内 有 股 闷热 之 气.
§ le nouveau-né s'agite dans son berceau. bien que le vieux ait laissé, pour entrer, ses sabots à la porte, son pas a fait craquer le plancher : l'enfant commence à geindre. la mère se penche hors de son lit, afin de le rassurer ; et le grand-père allume la lampe en tâtonnant, pour que le petit n'ait pas peur de la nuit. la flamme éclaire la figure rouge du vieux jean-michel, sa barbe blanche et rude, son air bourru et ses yeux vifs. il vient près du berceau. son manteau sent le mouillé ; il traîne en marchant ses gros chaussons bleus. louisa lui fait signe de ne pas s'approcher. elle est d'un blond presque blanc ; ses traits sont tirés ; sa douce figure mouton est marquée de taches de rousseur ; elle a des lèvres pâles et grosses, qui ne parviennent pas à se rejoindre et qui sourient avec timidité ; elle couve l'enfant des yeux – des yeux très bleus, très vagues, où la prunelle est un point tout petit, mais infiniment tendre.	§ 初生 的 婴儿 在 摇篮 里 扭动. 老人 进来 虽然 把 木靴 脱 在 门外, 走路 的 时候 地板 还是 格格 的 响 : 孩子 哼 啊 嗐 的 哭 了. 母亲 从 床上 探 出 身子 抚慰 他 ; 祖父 摸索 着 点 起 灯 来, 免得 孩子 在 黑夜 里 害怕. 灯光 照 出 老 约翰 · 米希尔 红红的 脸, 粗 硬 的 白 须, 忧郁 易怒 的 表情, 炯炯有神 的 眼睛. 他 走近 摇篮, 外套 发出 股 潮气, 脚下 拖 着 双 大 蓝布 鞋. 鲁 意莎 做 着 手势 叫 他 不要 走近. 她 的 淡黄 头发 差不多 象白 的 ; 绵羊 般 和善 的 脸 都 打皱 了, 颇 有些 雀斑 ; 没有 血色 的 厚 嘴唇 不大 容易 合拢, 笑 起来 非常 胆怯 ; 眼睛 很 蓝, 迷 迷惘 惘 的, 眼珠 只有 极 小 的 一 点, 可是 挺 温柔 ; -她 不胜 怜爱 的 瞅 着 孩子.
§ l'enfant s'éveille et pleure. son regard trouble s'agite. quelle épouvante ! les ténèbres, l'éclat brutal de la lampe, les hallucinations d'un cerveau à peine dégagé du chaos, la nuit étouffante et grouillante qui l'entoure, l'ombre sans fond d'où se détachent, comme des jets aveuglants de lumière, des sensations aiguës, des douleurs, des fantômes : ces figures énormes qui se penchent sur lui, ces yeux qui le pénètrent, qui s'enfoncent en lui, et qu'il ne comprend pas - il n'a pas la force de crier ; la terreur le cloue immobile, les yeux, la bouche ouverts, soufflant du fond de la gorge. sa grosse tête boursouflée se plisse de grimaces lamentables et grotesques ; la peau de sa figure et de ses mains est brune, violacée, avec des taches jaunâtres.	§ 孩子 醒 过来, 哭 了. 惊慌 的 眼睛 在 那儿 乱 转. 多 可怕 啊 ! 无边 的 黑暗, 剧烈 的 灯光, 浑沌 初 凿 的 头脑 里 的 幻觉, 包围 着 他 的 那 个 闷 人 的 、 蠕动 不已 的 黑夜, 还有 那 深不可测 的 阴影 中, 好似 耀眼 的 光线 一般 透 出来 的 尖锐 的 刺激, 痛苦, 和 幽灵, -使 他 莫名 片 妙 的 那些 巨大 的 脸 正 对 着 他, 眼睛 瞪 着 他, 直 透 到 他 心里 去 ... 他 没有 气力 叫喊, 吓 得 不能 动弹, 睁 着 眼睛, 张 着 嘴, 只 在 喉咙 里 喘气. 带 点 虚 肿 的 大胖 脸 扭 做 一 堆, 变成 可笑 而 又 可怜 的 怪 样子 ; 脸上 与 手上 的 皮肤 是 棕色 的, 暗红 的, 还 有些 黄 黄 的 斑点.

Pour rendre le sens vaguement péjoratif associé en français à *vieux vêtement*, il faut, en chinois, avoir recours à d'autres mots. La traduction mot à mot en chinois de : *vieux rideau* et *vieille caisse* ne signifierait pas forcément, que les objets considérés sont en mauvais état mais soulignerait simplement leur ancienneté, sans liaison explicite avec leur état au moment du récit. Fu Lei emploie 破 (po, *abîmé, déchiré*) et 破旧 (po jiu, *abîmé, usé, déchiré, etc.*) pour rendre accessible aux lecteurs chinois le sens original.

Tableau 7

Extrait de la concordance autour du pôle 老 (lao, *vieux*)

着 他的 要求 哼 一??歌词 没有 意义 的 **老调** 。 父亲 觉得 那种 音乐 是 胡闹 ； 可是 克利斯

那儿 摇晃 。 瘦削 的 树 好似 奇形怪状 的 **老人** 。 路旁 界石 上 的 反光 ， 象 青灰色 的

？， 尤其是 把 人家 的 敬意 看得 很重 的 **老人** 。 他们 常常 跟 他 说 些 过火 的 笑话 ，

而 一想 到 就 觉得 心灰意冷 。 # 可怜 的 **老人** ！ 在 无论 哪 方面 ， 他 都不能 完全 表露

党?他所 有的 小计 划 ， 仿佛 他们 俩 是 **老朋友** ； 他 说 他 怎样 想做 一个 象 哈斯莱 那样

?不会 说的 吧 ？ …… —— （ 他 指着 **老人** ） —— 瞧 ， 祖父 就在 那边 。 我 真爱

它们 象 牛 ， 象 巨人 ， 象 帽子 ， 象 **老婆婆** ， 象 广漠 无垠 的 风景 。 他 和 它们 低声

忱?， 快活 得 脸红 了 。 比他 更快 活的 **老人** ， 装着 若无其事 的 声音 和 他 说 （ 因为

器具 和 动物 的 尸身 ， 裹着 大氅 ， 象 **老太太** 般 ， 一边 庄严 的 前进 ， 一边 行 着 礼

低 的 吼 着 。 孩子 一个又一个 的 听 上 **老** 半天 ， 听 它们 低下 去 ， 没有 了 ； 它们

的时候 。 往往 你得 不 声 不 响的 等 个 **老** 半天 ， 正当 克利斯 朵 夫 想着 “ 他 今晚

但 就 因为 厌恶 ， 反而 常常 要看 。 他 **老** 半天 的 瞪着 它们 ， 不时 向 四下里 溜 一眼

贵族 ??生 的 家长 出来 散步 。 那时 他 **会老** 半天 的 停下来 ， 深深 的 鞠躬 ， 说着 一大

： # “ 噢 ！ 祖父 ！ 祖父 ！ …… ” # **老人** 把 他 拉到 身边 。 他 扑在 老人 膝上 ，

峡?罢 。 ” # “ 那 也 该 回来 啦 ， “ **老人** 不高兴 的 说 。 # 他 踌躇 了 一会 ， 很不

命运 。 他 尤其 为 一个 美人 儿 颠倒 ， **不老** 不少 的 年纪 ， 金黄 的 长发 ， 大得 有点

所 教 的 东西 了 。 给 骂 了 一顿 ， 他 **老大** 不愿意 的 继续 下去 。 这样 当然 招来 了

， 他 没有 ， ” 鲁 意 莎 抢着 回答 。 # **老人** 瞅着 她 ， 她 把 眼睛 躲开 了 。 # “ 哼

发愁 ， 时时刻刻 从 窗 里 张望 。 终于 **老人** 出现 了 ， 他们 俩 动身 了 。 他的 心 在

似的 告诉 他 ， 说 有些 东西 给 他 看 。 **老人** 打开 书桌 ， 检出 一本 乐器 放在 钢琴 上

岁 的 姑娘 ， 腮 帮 通红 ， 非常 壮健 ， **老** 带着 笑容 。 奥蒂 丽 的 长处 正好 和 克拉拉

十八世纪 的 雕 有 人像 的 柜子 ； 那 是 **老人** 从来不 肯 割爱 的 ， 虽然 古董 商 华姆塞

问道 ： # “ 那末 您 呢 ， 祖父 ？ ” # **老人** 打了 个 寒噤 。 # “ 什么 ？ “ 他 问 。

有心 装做 对 故事 的 下文 满不在乎 ， 使 **老人** 大为 难过 。 —— 但 眼前 他 是 完全 给

” # 孩子 迷 迷 忽忽 的 ， 对着 灯光 和 **老人** 的 目光 愣住 了 ， 这 时 才 醒过来 ， 哭了

也 做 过 这些 东西 ？ ” # “ 当然 ， ” **老人** 的 声音 有点儿 不高兴 。 # 说完 他 不做

时候 常常 带着 他 一块儿 去 。 孩子 拉着 **老人** 的 手 在旁边 急急忙忙 的 搬 着 小步 。 他们

夜里 ， 还能 看出 他 憔悴 的 脸 ， 好似 **老人** 的 一样 。 她 开始 ??睡 了 ， 乱哄哄 的

吃了一惊 。 大家 一起 笑了 ； 大公 爵 向 **老人** 道贺 ， 他却 慌 做 一团 ， 想 解释 又 解释

， 影子 的 头 会 爬上 去 ， 过后 又 回到 **老** 地方 ； 口 环 变得 很大 ， 象个 破 气球

口 ， 他 茫然若??， 发觉 自己 还是 在 **老** 地方 ， 在 黑 ？？的 楼梯 上 。 在 几步

的 某一个 人 ， 但 英勇 的 事迹 使 他 和 **老人** 都 骄傲 得 心花怒放 ， 仿佛 那些 事 就是

朵 夫 立刻 凑 上去 。 他们 俩 很 投机 。 **老人** 非常 喜欢 孙子 ； 有个 愿意 听 他 说话 的

嚷 起来 。 母亲 嘲笑 他 。 曼希沃 说是 **老人家** 疯了 ， 与其 把 孩子 弄得 神魂颠倒 ， 还不

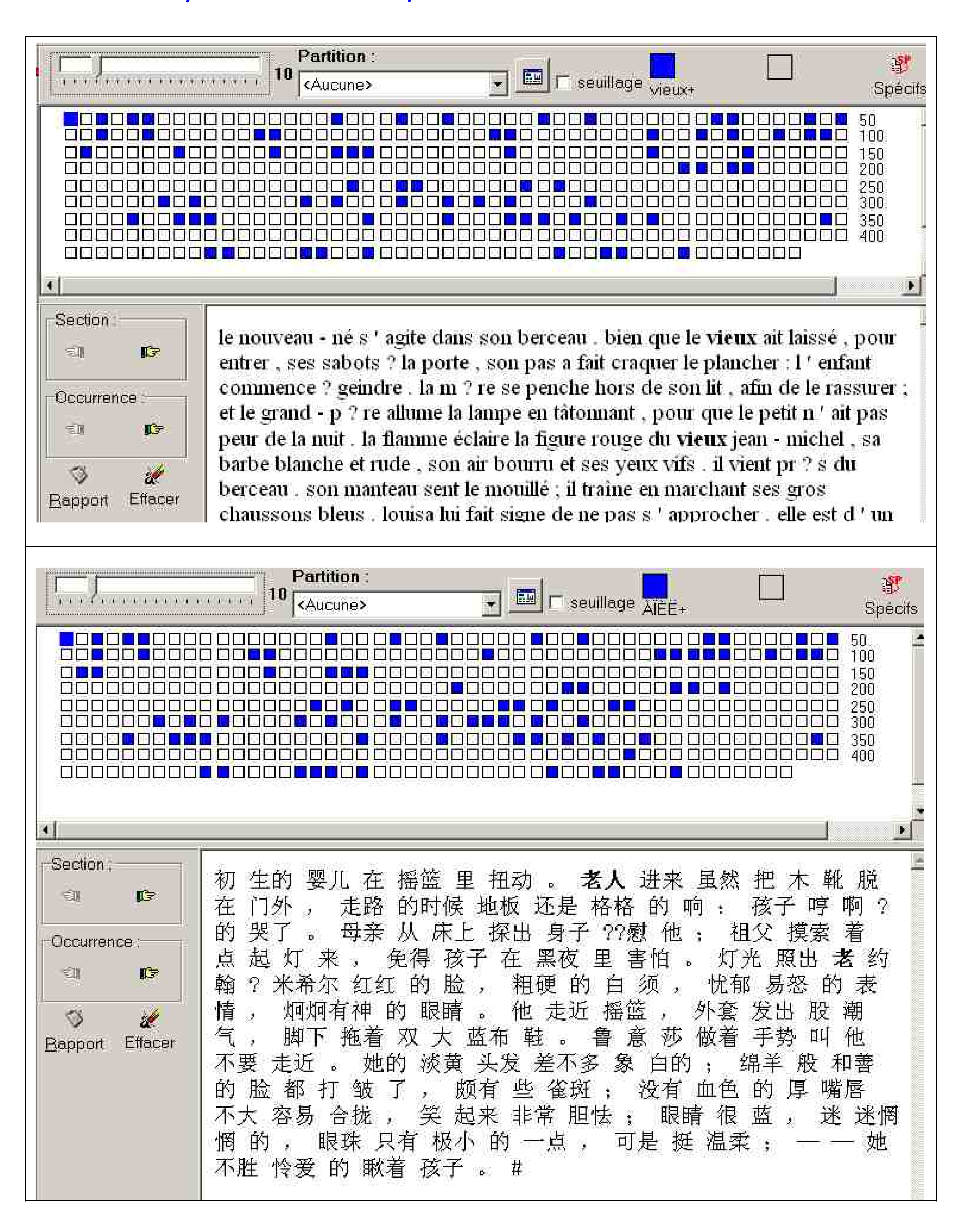


Figure 8 :

Localisation des correspondances de *vieux* et 老 dans le bitexte

à l’aide du logiciel *Lexico3*.

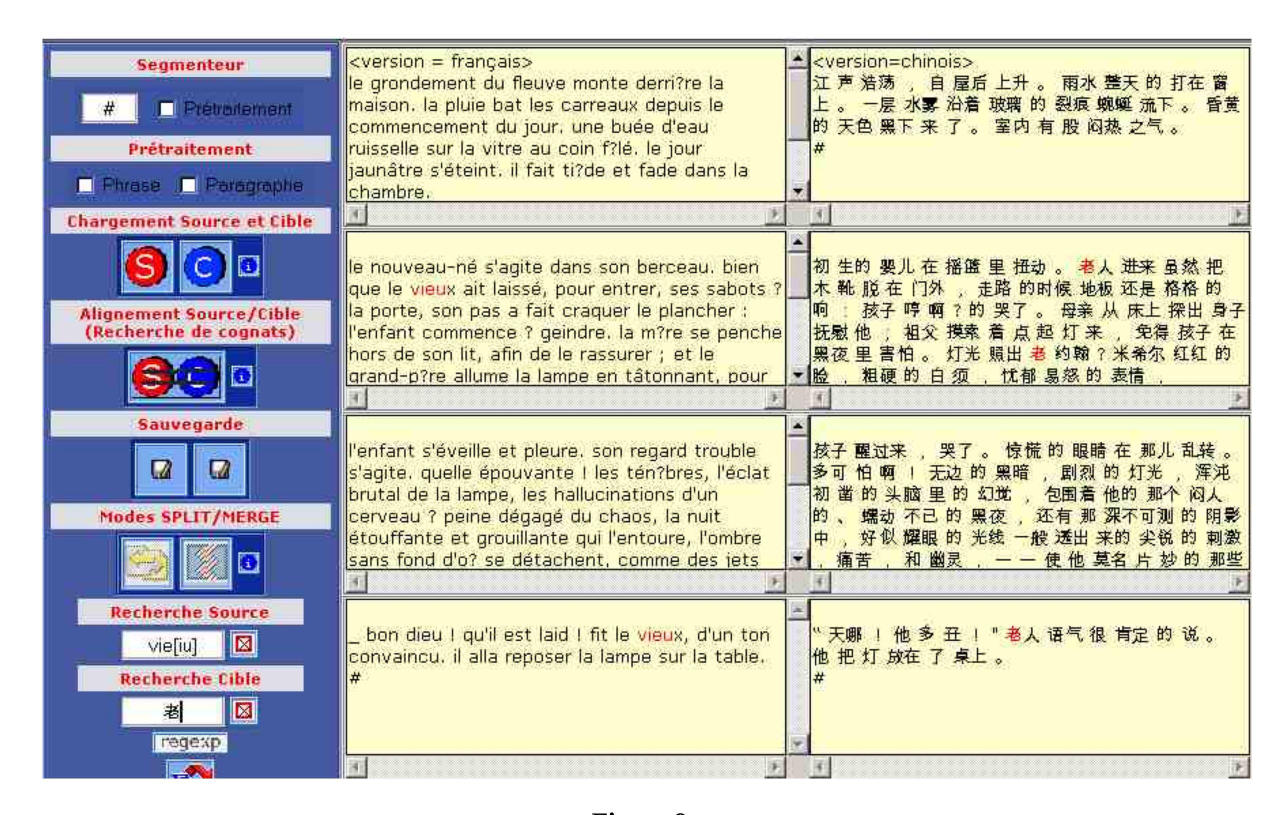


Figure 9 :

Visualisation des correspondances de *vieux* et 老 dans le bitexte
à l'aide du logiciel mkAlign.

La localisation des concordances et des discordances dans la localisation des termes qui qui sont réputés constituer des équivalences traductionnelles permet d'approfondir l'étude traductologique et de mieux cerner les techniques propres à chaque traducteur pour rendre compte du sens véhiculé par le texte source.

Tableau 8 :

Traductions attestées dans le volet chinois pour le terme *vieux*

français	traduction chinoise	français	traduction chinoise
vieille maison	旧屋子 (maison ancienne)	de vieux amis	老朋友 (vieux amis)
vieille ficelle	旧绳子 (ficelle usagée)	vieux grand père	祖父 (grand-père)
de vieux habits	旧衣衫 (vêtements usagés)	vieux grand père	老祖父(vieux grand-père)
vieux veston bleu	旧蓝上装 (veston usagé)	le vieux jean-michel	老祖父(vieux grand-père)
vieille chanson	老调(mélodie ancienne)	le vieux	老人家 (un vieil homme)
vieille chanson	老歌 (chanson ancienne)	pauvre vieux	老人家 (vieil homme)
viel escalier	黑魆魆的楼梯(escalier noir)	vieilles dames	老太太 (vieilles dames)
vieux rideau	破帘子 (rideau usagé)	vieilles dames	老婆婆 (vieilles dames)
vieille caisse	破旧匣子 (caisse abîmée)	il vieillissait	年纪越大(il prenait de l'âge)

7 Conclusion

La complexité apparente, le système d'écriture chinois ne constitue pas un obstacle incontournable à l'exploration textométrique des textes. Les traitements informatisés élaborés pour les textes codés à l'aide d'écritures alphabétiques peuvent être adaptés, moyennant des modifications mineures à l'étude des textes chinois.

Malgré des difficultés importantes dans la définition de l'entité *mot* en chinois, l'introduction de cette notion et sa prise en charge par des logiciels de segmentation automatique permet d'augmenter l'efficacité de l'exploration textométrique du bitexte franco-chinois et de dépasser l'exploration fondée sur les caractères *hanzis* considérés comme des entités isolées.

Les résultats, obtenus sur la base de la comparaison textométrique du bitexte aligné découpé en mots ouvrent, au plan traductologique, des pistes de comparaison qui semblent extrêmement prometteuses. Elles permettent d'envisager la comparaison simultanée des moyens lexicaux utilisés dans les corpus de traduction mis en confrontation et des procédés employés par les traducteurs pour faire saisir à leurs lecteurs les différents sens, nuances et connotations véhiculés par le texte d'origine.

8 Références

ALLETON V. 1997. *L'écriture chinoise*, « Que sais-je ?», 5è édition corrigée, 1re édition : 1970, Paris, Presses universitaires de France.

FU LEI (傅 雷). 1998 .*La grande série de la traduction de Fu Lei 傅雷译文全集* , He fei, Éditions de l'art d'An Hui, 安徽文艺出版社.

FLEURY S., MKAlign : *Manuel d'utilisation*, http://tal.univ-paris3.fr/mkAlign/mkAlignDOC.htm

GRANGER S., LEROT J., PETCH-TYSON S. (eds.). 2003. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Amsterdam – New York, Editions Rodopi.

HABERT B., NAZARENKO A., et SALEM A. 1997. *Les linguistiques de corpus.* Paris, Armand Colin/Masson.

HOA M. 2005. *C'est du chinois*! I, volume "Lire et écrire", 3è édition. Paris, Édition You-Feng.

LEBART L., SALEM A., *Statistique textuelle*, Paris, Dunod, 1994, téléchargeable sur le site : http://www.cavi.univ-paris3.fr/lexicometrica/livre/st94/st94-tdm.html

OLOHAN M. 2004. *Introducing Corpora in Translation Studies*. London and New York, Routledge.

SALEM A., "Introduction à la résonance textuelle", Actes *des 7èmes Journées d'analyse des données textuelles,* Louvain la neuve, 2004

WEI N. et alii. 2005. *Corpora in use 语料库应用研究.* Shanghai, Éditions de l'enseignement des langues étrangères de Shanghai 上海外语教育出版社.

ZIMINA, M. 2004. *Approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*. Thèse de doctorat, Université de la Sorbonne nouvelle – Paris3.

ZIMINA, M. 2005. *Topographie bi-textuelle et approches quantitatives de l'extraction de ressources traductionnelles à partir de corpus parallèles*. Actes des 7es Journées scientifiques du Réseau de chercheurs "Lexicologie, Terminologie, Traduction", Institut supérieur des traducteurs et interprètes (ISTI), Bruxelles.

ZIPF, G., K. 1935. *The Psychobiology of Language, an introduction to Dynamic Philology.* Boston, Houghton-Mifflin.

ZHOU Q., DUAN H., 周 强, 段慧明. 2007. *Traitement de segmentation et de marquage des mots dans les corpus chinois modernes 现代汉语语料库加工中的切词与词性标注处理,* disponible sur http://hi.baidu.com/jagard/blog/item/dcdb653844fd842097ddd8ec.html

9 Fonctionnalités *Lexico3* utilisées dans cette exploration

N°	*Fonctionnalité*	*Résultat*
5.5	Courbe d'accroissement des caractères (hanzis)	*Figure 5*
5	Principales caractériques lexicométriques (PCLC)	*Tableau 4*
5.5	Accroissement du vocabulaire (chinois et français)	*Figure 6*
5.4	Diagramme de Pareto (chinois et français)	*Figure 7*
7	Carte des sections (volets français et chinois)	*Figure 8*